بخش اول) راه‌اندازی خوشه Hadoop

نقش‌های گره‌های خوشه:

```
h-user@hadoopn1:~$ jps
46292 Jps
3675 NameNode
5471 ResourceManager
3919 SecondaryNameNode
```

```
h-user@hadoopn2:~$ jps
19093 NodeManager
19752 Jps
18140 DataNode
```

```
h-user@hadoopn3:~$ jps
36112 Jps
16848 NodeManager
15031 DataNode
```

گره‌ی Primary نقش‌های NameNode و SecondaryNameNode را برای مدیریت خوشه HDFS و نقش ResourceManager را برای مدیریت منابع در Yarn بر عهده دارد.

گره‌های Secondary نقش DataNode برای ذخیره فایل‌های HDFS و NodeManager برای ارتباط با گره Primary در راستای پذیرش و مدیریت job ها را بر عهده دارند.

Datanode ها:

Configured Capacity:	38.13 GB
Configured Remote Capacity:	0 B
DFS Used:	56 KB (0%)
Non DFS Used:	30.63 GB
DFS Remaining:	5.52 GB (14.47%)

In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
hadoopn2:9866 (192.168.1.108:9866)	http://hadoopn2:9864	0s	51m	19.07 GB	0	28 KB (0%)	3.2.2
hadoopn3:9866 (192.168.1.109:9866)	http://hadoopn3:9864	1s	51m	19.07 GB	0	28 KB (0%)	3.2.2

به هر گره 20GB حافظه دیسک اختصاص داده شده است که بیشتر آن توسط سیستم عامل و برنامه‌ها اشغال شده و از هر گره حدود 2.26GB حافظه برای HDFS باقی مانده است.

گره‌های Yarn:

در کانفیگ yarn-site.xml برای گره‌های Secondary، منابع زیر نوشته شد:

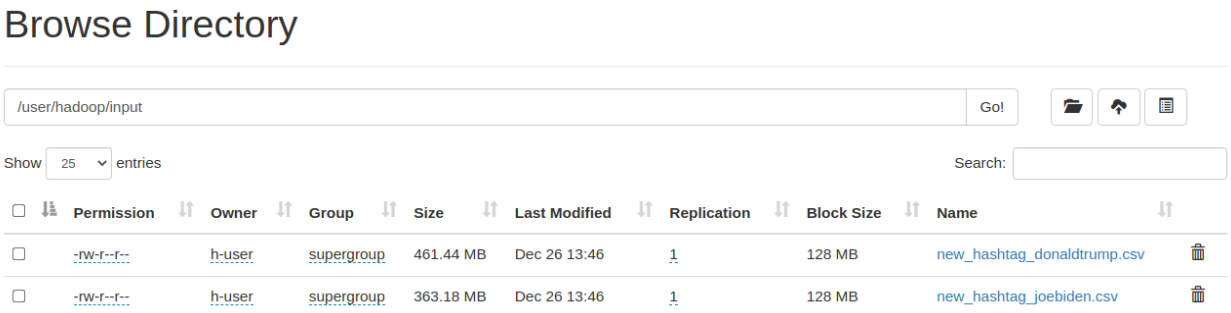
```
<property>
  <name>yarn.nodemanager.resource.cpu-vcores</name>
  <value>2</value>
</property>


<property>
  <name>yarn.nodemanager.resource.memory-mb</name>
  <value>2048</value>
</property>
```

در نتیجه در WebGUI مربوط به Yarn هم همین منابع را خواهیم دید:

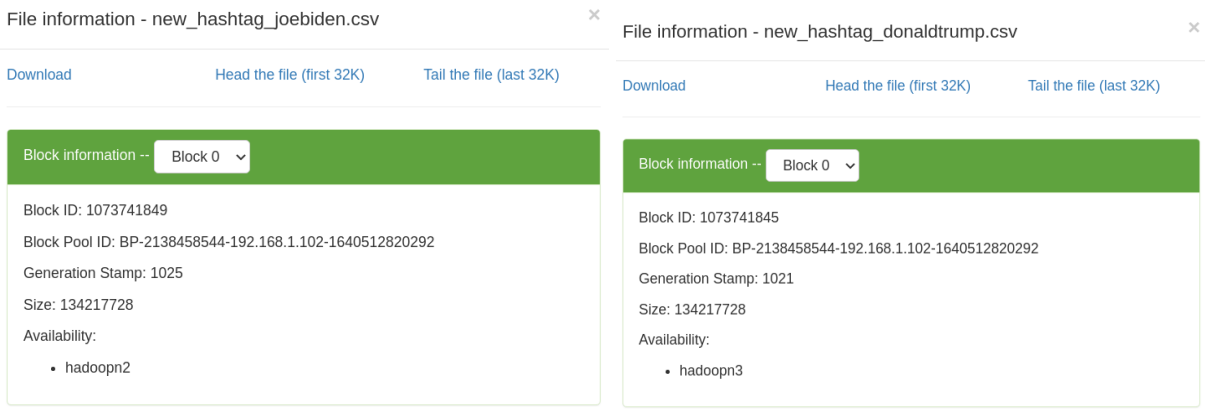
Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %	VCores Used	VCores Avail	Phys VCores Used %
/default-rack	RUNNING	hadoopn3:32933	hadoopn3:8042	Sun Jan 09 18:34:55 +0330 2022		0		0 B	2 GB	38	0	2	0
/default-rack	RUNNING	hadoopn2:40281	hadoopn2:8042	Sun Jan 09 18:34:55 +0330 2022		0		0 B	2 GB	38	0	2	0

بخش دوم) توسعه و اجرای برنامه‌های MapReduce

قرار دادن فایل‌های دیتاست درون HDFS:



یک فایل در گره hadoopn2 و دیگری در hadoopn3 قرار دارد:



مسئله اول - شمارش تعداد لایک‌ها و ریتوییت‌ها

Map: گرفتن یک خط (یک توییت) از فایل‌های CSV، جستجو برای هشتگ‌های مربوط به دو کاندید، تولید یک tuple که کلید آن ۳ حالت دارد (یا Trump است، یا Biden یا Both) و مقدار آن یک متن، که تعداد لایک‌ها و ریتوییت‌های آن توییت است.

Reduce: محاسبه مجموع لایک‌ها و ریتوییت‌ها برای هر یک از ۳ حالت با for زدن روی آن‌ها

نکته: از آن‌جا که توییت‌های یک فایل CSV که مربوط به هر دو کاندید می‌باشند، عینا در فایل CSV دیگر هم تکرار شده‌اند، لازم است که پس شمارش آن‌ها، در نهایت تعداد لایک و ریتوییت‌ها را تقسیم بر ۲ کنیم.

اجرا:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime
application_1641740697487_0001	h-user	LR Count	MAPREDUCE	default	0	Sun Jan 9 18:51:36 +0350 2022	Sun Jan 9 18:51:37 +0350 2022

Task Type	Progress	Total	Pending	Running	Complete
Map		7	5	1	1
Reduce		1	1	0	0

خروجی:

```
h-user@hadoopn1:/home/hadoopn1/MR$ hadoop fs -cat /user/hadoop/output/part-r-00000
Both    1156293 245235
Donald Trump    4920678 1159176
Joe Biden       5606807 1164631
```

مسئله دوم - شمارش توییت‌های مربوط به دو کاندیدا در کشورهای مختلف

Map: گرفتن یک یک توییت از فایل‌های CSV، جستجو برای هشتگ‌های مربوط به دو کاندید، جستجو در فیلد country برای یافتن کشور، ایجاد یک tuple که کلید آن نام کشور است و مقدار آن ۳ حالت دارد (0: توییت مربوط به Biden است، 1: توییت مربوط به Trump است، 3: توییت مربوط به هر دو نامزد است)

Reduce: محاسبه مجموع توییت‌های مربوط به نامزدها در هر کشور و محاسبه درصدها

خروجی:

```
h-user@hadoopn1:/home/hadoopn1/MR$ hadoop fs -cat /user/hadoop/output/part-r-00000
america 0.17209664 0.36693755 0.4609658 965167
austria 0.20892581 0.24834576 0.54258764 7103
canada  0.17639822 0.2993653 0.5242365 82087
emirates        0.23583907 0.3771837 0.3868449 7556
france  0.22986715 0.2985417 0.47159114 115545
germany 0.2105401 0.25931028 0.5301412 118310
iran    0.21224944 0.28017816 0.5075724 4490
italy   0.21502227 0.3387626 0.44621515 68272
mexico  0.22791281 0.35333645 0.41875073 34228
netherlands     0.20340343 0.31384596 0.48275062 47364
spain   0.22440115 0.29297403 0.48258266 23712
```

مسئله سوم - شمارش توییت‌های مربوط به دو کاندیدا با توجه به موقعیت جغرافیایی

MapReduce: مشابه مسئله قبل با این تفاوت که کشور کاربر با توجه به فیلدها lat و long بدست می‌آید.

خروجی:

```
h-user@hadoopn1:/home/hadoopn1/MR$ hadoop fs -cat /user/hadoop/output/part-r-00000
america 0.1161915 0.32405332 0.5597552 98527
france  0.14056997 0.31123918 0.54819083 12492
```

تفاوت نتایج با مسئله ۲ می‌تواند به این دلیل باشد که بعضی از کاربران ممکن است نام کشور را در اکانت توییر خود را وارد نکرده باشند (یا اشتباه وارد کرده باشند) یا از VPN استفاده کنند، در نتیجه موقعیت جغرافیایی محاسبه شده از Lat و Long مربوط به IP شان، با موقعیت موجود در فیلد country متفاوت خواهد بود.